Technical Note: WHD Mixing with Two-Path ReLU

$S = W_1HD + W_2H(I - D)$

---

1. Core Idea

We extend the two-path ReLU construction:

$S = W_1D + W_2(I - D)$

by inserting a fast mixing transform:

$S = W_1HD + W_2H(I - D)$

where:

- H is the Walsh–Hadamard transform
- D encodes ReLU decisions

---

2. The Decision Matrix (D)

The matrix (D) is best understood as a **conceptual tool**:

- D is a diagonal matrix
- Each diagonal entry is:

$$D_{ii} = \begin{cases} 1 & \text{if the } i\text{-th pre-activation is positive} \\ 0 & \text{otherwise} \end{cases}$$

So:

- D selects the “positive activation” pathway
- I - D selects the complementary pathway

Importantly:

- D is **implicit** in standard implementations
- It arises from ReLU decisions and **does not need to be explicitly constructed**
- It is introduced to reason about the system as a composition of linear operators

---

3. Layer Interpretation

For input x:

1. **Mix**

   $z = Hx$

2. **Gate (ReLU decisions)**

   $D = \text{diag}(1_{z>0})$

3. **Two-path linear combination**

   $y = W_1Dz + W_2(I - D)z$

Equivalently:

$y = (W_1HD + W_2H(I - D))\ x$

---

4. Why Insert H?

Without H:

- Gating is **axis-aligned**
- Submodels are highly correlated
- Weight usage is redundant

With H:

- Inputs are **globally mixed before gating**
- Decisions depend on **distributed patterns**, not individual coordinates
- Submodels become **less correlated**

This reduces unwanted structure in how weights are selected.

---

5. Effect on the Induced System Family

Each realization of D defines a linear operator:

$S(D) = W_1HD + W_2H(I - D)$

Key change:

- Without H: partitions of space are coordinate-aligned
- With H: partitions are effectively **rotated / mixed**

Result:

- More diverse linear regions
- Less overlap between regions
- Better parameter efficiency

---

6. Correlation Reduction Mechanism

The main benefit comes from **decorrelation before gating**:

- H spreads each input component across all coordinates
- Each gating decision depends on many input dimensions
- Weight matrices W_1, W_2 are applied to **less correlated inputs**

This reduces:

- Repeated use of similar features
- Degenerate or redundant submodels

---

7. Stacking Behavior

For multiple layers:

$$x_{l+1} = W_1^{(l)} H D^{(l)} x_l + W_2^{(l)} H (I - D^{(l)}) x_l$$

Properties:

- Each layer remixes the representation
- Gating decisions evolve with depth
- The network behaves like a **deep ensemble of decorrelated linear systems**

---

8. Practical Notes

- H is fixed, fast, and parameter-free
- Dimensions are typically powers of two (padding if needed)
- W_1 and W_2 can be initialized similarly to start near linear behavior
- The implicit D structure is sufficient for implementation

---

9. Conceptual Summary

The key shift is:

> Apply gating in a **mixed coordinate system** rather than the original basis.

This turns ReLU from:

- A coordinate-wise selector

into:

- A selector over **global, distributed patterns**

leading to:

- Reduced correlation in weight usage
- More expressive partitioning of input space
- Improved efficiency of the induced piecewise linear system